Keyword Weight Propagation for Indexing Structured Web Content

Jong Wook Kim, and K. Selcuk Candan

Comp. Sci. and Eng. Dept
Arizona State University
{jong, candan}@asu.edu

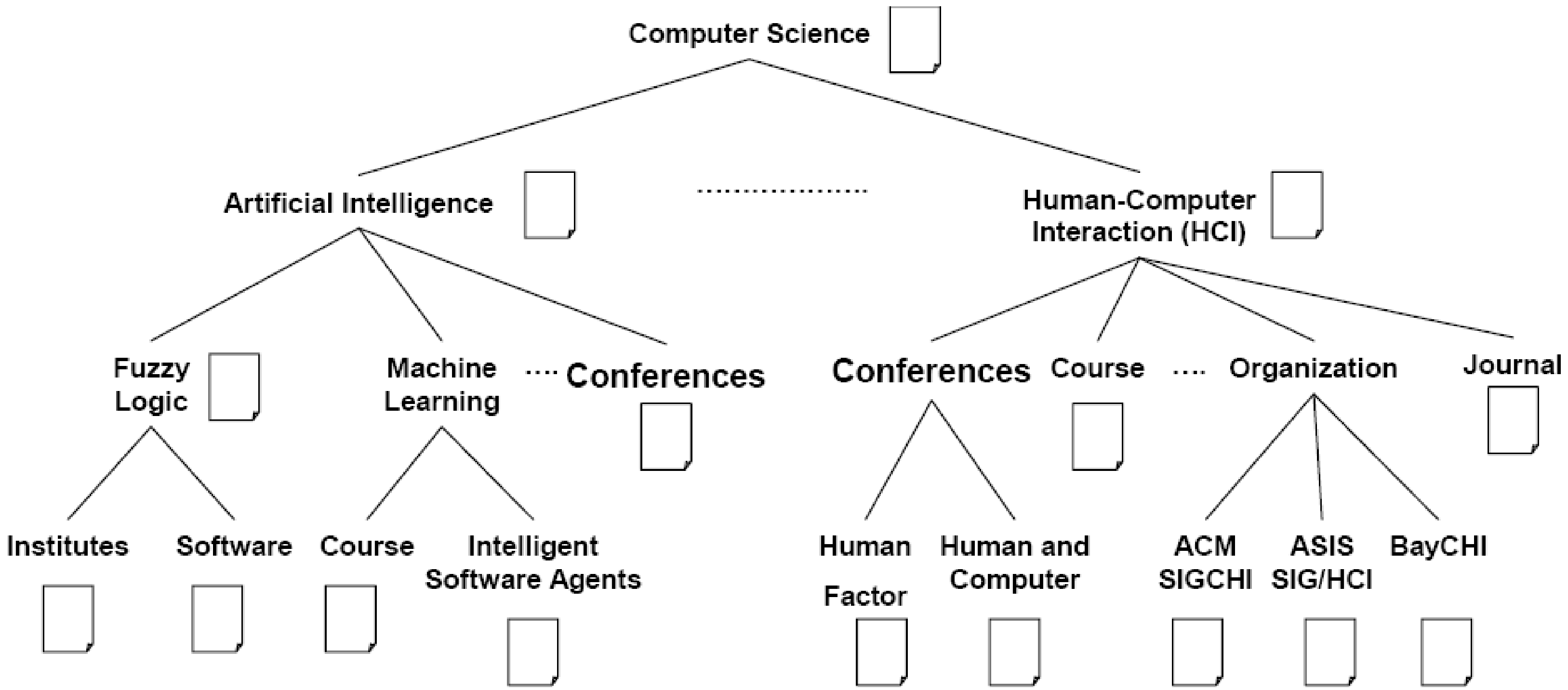
WebKDD 2006 Workshop on Knowledge Discovery on the Web, Aug. 20, 2006,
Philadelphia, PA, USA

Table

- ➢ Motivation
- ❑ Approach
- ❑ Related Work
- ❑ Relative Content of Entries
- ❑ Keyword Propagation
  - ❑ Keyword Propagation between a Pair of Entries
  - ❑ Keyword Propagation across a Complex Structure
- ❑ Experiment
- ❑ Conclusion and Future Work

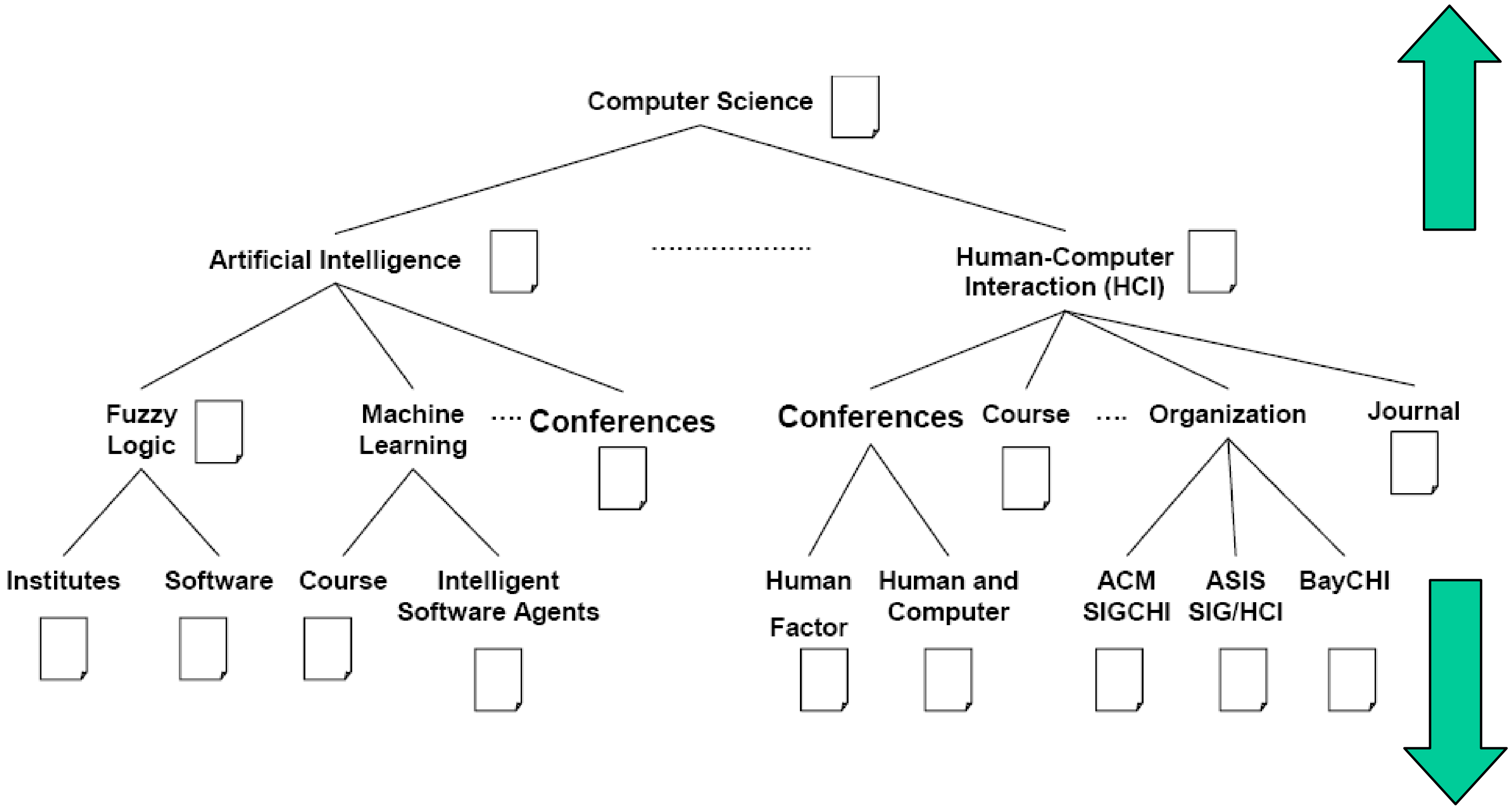
Motivation

- Many web sites and portals organize content in a navigation hierarchy



Motivation

- Many web sites and portals organize content in a navigation hierarchy



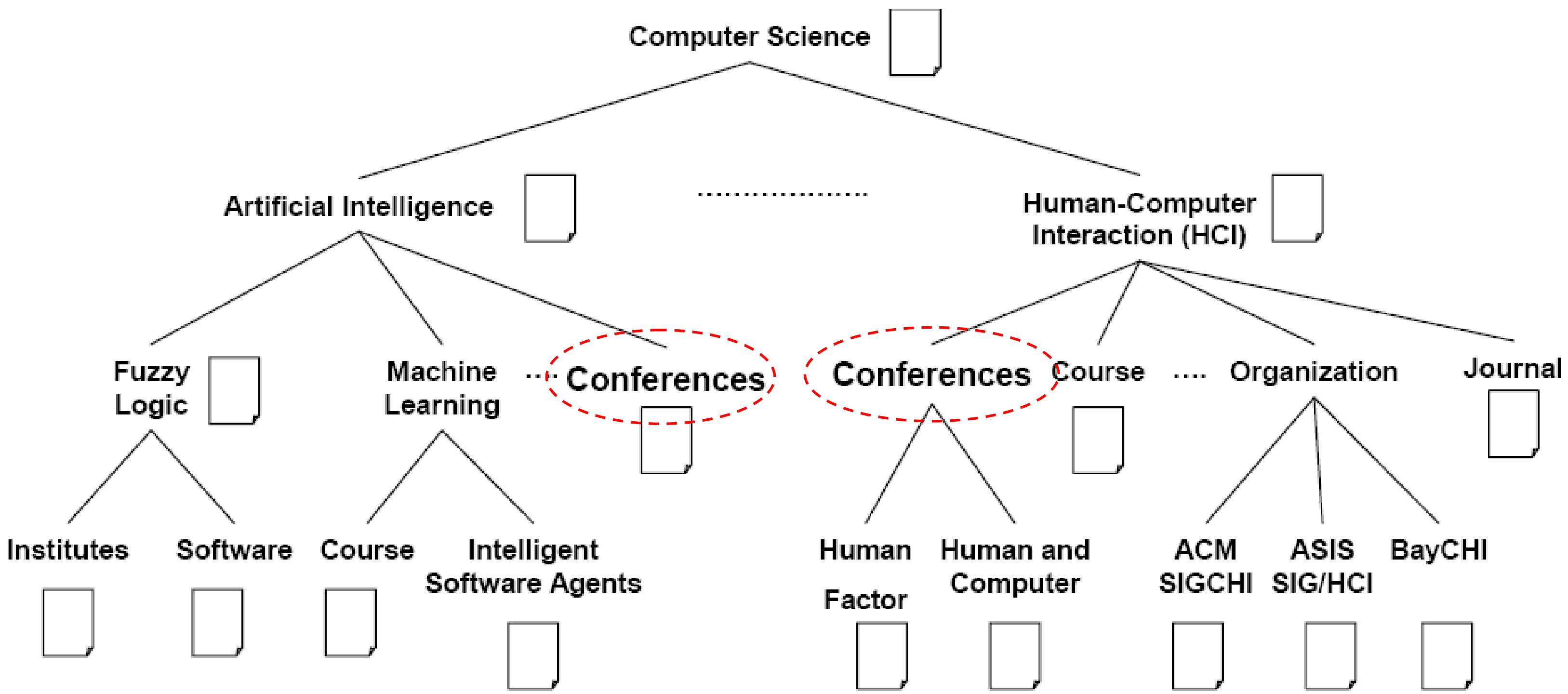
Motivation

- Many web sites and portals organize content in a navigation hierarchy

- A navigation hierarchy
  - Effective when browsing to find a specific content
  - Semantic relationships between the data contents
    - ☆ Generalization/ Specialization

Motivation

- Keyword contents of the intermediate nodes may describe their content in the hierarchy ambiguously



The Yahoo CS hierarchy

Motivation

- In a navigational hierarchy, keyword searches are usually directed
  - to the root of the hierarchy, or
    - ✰ Undesirable topic drift
  - to the leaves
    - ✰ May not be enough to satisfy the query
- It is important for individual nodes to be properly indexed

Table

Approach

- Keyword and keyword weight propagation
  - Enrich the individual nodes with the contents of the neighboring nodes
- How to decide what to propagate and how much?
  - The original semantic structure should be preserved
    - ☆ Generalization/ Specialization
- Challenge
  - How to represent the semantic structure (i.e., generalization/ specialization) between nodes?
  - How to determine the degree of keyword inheritance?

Approach

- Contributions of the Paper
  - Develop a method for discovering and quantifying the generalization/ specialization relationship between entries in a navigation hierarchy
  - Develop a keyword propagation algorithm using this relationship

Table

- ❑ Motivation
- ❑ Approach
- ➢ Related Work
- ❑ Relative Content of Entries
- ❑ Keyword Propagation
  - ❑ **Keyword Propagation between a Pair of Entries**
  - ❑ **Keyword Propagation across a Complex Structure**
- ❑ Experiment
- ❑ Conclusion and Future Work

Related Work

- Score and Keyword Frequency Propagation
  - Propagate the relevance score [Shakery, and Zhai, TREC'03]
  - Propagate the term frequency value [Savoy et al. JASIS'97] [Song et al. TREC'04]
  - Propagate the relevance score and the term frequency value [Qin et al. SIGIR'05]

Table

- Motivation
- Related Work
- Approach
- Relative Content of Entries
- Keyword Propagation
  - Keyword Propagation between a Pair of Entries
  - Keyword Propagation across a Complex Structure
- Experiment
- Conclusion and Future Work

Relative Content of Entries

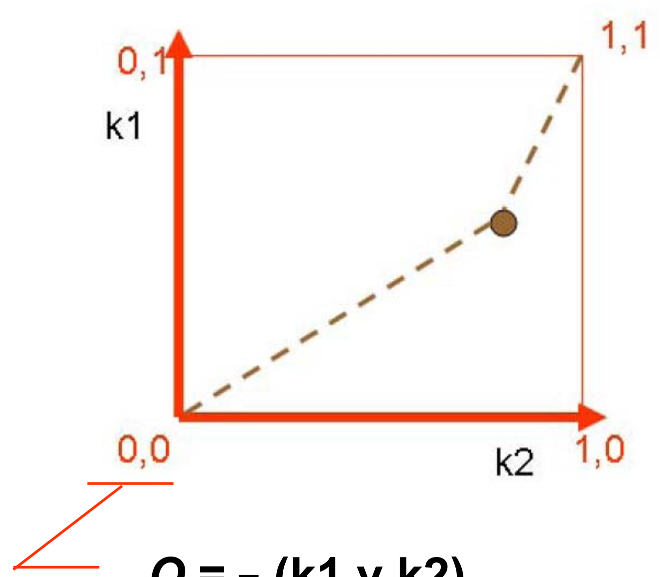
- In a navigation hierarchy,
  - A specialized entry corresponds to more constrained concept
    - ✩ As one moves down in a hierarchy, the nodes get more specialized
  - A general entry is less constrained
    - ✩ As one moves up in a hierarchy, the nodes get more generalized.

Relative Content of Entries

- Intuition
  - Given two entries, A and B (A is an ancestor of B),
    - ☆ Assume
      - **A has three keyword (k1, k2, k3) , and**
      - **B has two keyword (k2, k3)**
    - ☆ "Entry A is more general than B" → A being less constrained than B by keywords
    - ☆ If B is interpreted as k2 v k3, then A should be interpreted as k1 v k2 v k3
      - Less constrained than k2 v k3
    - ☆ Interpreted as the disjunction of keywords

Relative Content of Entries

- In extended boolean model [Salton 83],
  - OR-ness
    - ✰ An entry further away from ***O*** better matches the k1 v k2
    - ✰ Measured as a distance from ***O***



$O = \neg$ (k1 v k2)

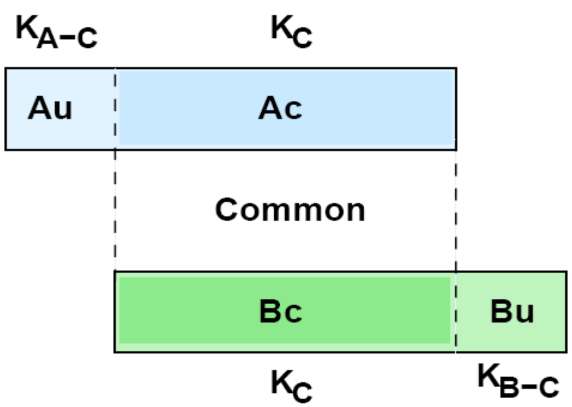
Relative Content of Entries

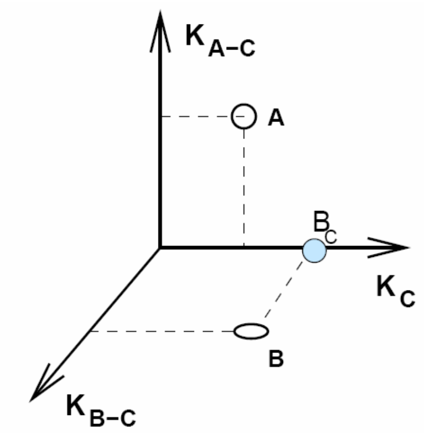
- Given two entries, A and B (A is an ancestor of B),
  - Assume
    - ☆ A has three keyword (k1, k2, k3) , and
    - ☆ B has two keyword (k2, k3)
  - How much entry A and B represent a disjunct ?
    - ☆ $|\vec{A}-\vec{O}|=|\vec{A}|$, $|\vec{B}-\vec{O}|=|\vec{B}|$
  - If A is more general than B, then

$$|\vec{A}-\vec{O}|>|\vec{B}-\vec{O}|$$

Relative Content of Entries

❑ Visual representation of the keyword contents





❑ Relative Content

$$R_{AB} = \frac{|\overrightarrow{A}|}{|\overrightarrow{B_C}|} = \frac{|\overrightarrow{A_U} + \overrightarrow{A_C}|}{|\overrightarrow{B_C}|}$$

Measure whether the additional keywords (A_U) make A more general or less general than B_C

Table

- Motivation
- Approach
- Related Work
- Relative Content of Entries
- Keyword Propagation
  - Keyword Propagation between a Pair of Entries
  - Keyword Propagation across a Complex Structure
- Experiment
- Conclusion and Future Work

Keyword Propagation between a pair of entries

- The purpose of keyword propagation
  - Enrich the entries in a navigational hierarchy
  - The original semantic properties (i.e., relative generality) should be preserved
- Propagation Degree, α
  - Govern how much keyword weights two neighboring entries should exchange

Keyword Propagation between a pair of entries

- **Propagation Degree, α**
  - Given two entries, *A* and *B*,
    - ☆ a_i : weight associated with keywords $k_i \in K_A$
    - ☆ b_i : weight associated with keywords $k_i \in K_B$
  - *A'* and *B'*
    - ☆ Enriched entries after keyword propagation
  - For all $k_i \in K_{A'}$
    - ☆ If $k_i \in (K_A - K_B)$, then $a'_i = a_i$
    - ☆ If $k_i \in (K_A \cap K_B)$, then $a'_i = a_i + \alpha b_i$
    - ☆ If $k_i \in (K_B - K_A)$, then $a'_i = \alpha b_i$
  - For all $k_i \in K_{B'}$
    - ☆ If $k_i \in (K_A - K_B)$, then $b'_i = \alpha a_i$
    - ☆ If $k_i \in (K_A \cap K_B)$, then $b'_i = b_i + \alpha a_i$
    - ☆ If $k_i \in (K_B - K_A)$, then $b'_i = b_i$

Keyword Propagation between a pair of entries

- Propagation Degree, α
  - *A'* and *B'* are located in a common keyword space
    - ☆ $K_C = K_{A'} = K_{B'} = K_A \cup K_B$
  - After keyword propagation, relative content should be preserved

$$R_{A'B'} = R_{AB}$$

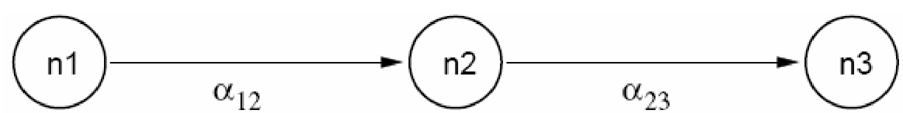
$$\Rightarrow \quad R_{A'B'} = \frac{|\vec{A}|}{|\vec{B_C}|} = \frac{|\vec{A'}|}{|\vec{B'}|} = R_{AB}$$

Table

- Motivation
- Approach
- Related Work
- Relative Content of Entries
- Keyword Propagation
  - Keyword Propagation between a Pair of Entries
  - Keyword Propagation across a Complex Structure
- Experiment
- Conclusion and Future Work

Keyword Propagation across a Complex Structure

- Let *H(N,E)* be a navigation hierarchy,
  - *N* : the set of nodes
  - *E* : the set of edges
- Propagation Adjacency Matrix, *M*
  - If there is an edge $e_{ij} \in E$, then both *(i,j)* and *(j,i)* of *M* are equals to α_{ij} (the pairwise propagation degree)
  - Otherwise, both *(i,j)* and *(j,i)* of *M* are equal to 0.



$$\begin{pmatrix} 0 & \alpha_{12} & 0 \\ \alpha_{12} & 0 & \alpha_{23} \\ 0 & \alpha_{23} & 0 \end{pmatrix}$$

Keyword Propagation across a Complex Structure

- ❑ Keyword Propagation Process
  - ❑ Given a hierarchy, *H(N,E)*
    - ☆ *T* : Term-node matrix
    - ☆ *M* : Propagation Adjacency matrix
  - ❑ Term Propagation Matrix
    - ☆ $P = T M$

$$\underset{\textbf{P}}{\begin{pmatrix} 0 & \alpha_{12}K1 & 0 \\ \alpha_{12}K2 & \alpha_{12}K2 & \alpha_{23}K2 \\ \alpha_{12}K3 & \alpha_{23}K3 & \alpha_{23}K3 \end{pmatrix}} = \begin{array}{cc} & \textbf{node} \\ \textbf{term} & \underset{\textbf{T}}{\begin{pmatrix} K1 & 0 & 0 \\ K2 & K2 & 0 \\ 0 & K3 & K3 \end{pmatrix}} \end{array} \underset{\textbf{M}}{\begin{pmatrix} 0 & \alpha_{12} & 0 \\ \alpha_{12} & 0 & \alpha_{23} \\ 0 & \alpha_{23} & 0 \end{pmatrix}}$$

Inherited from its neighbors in *M*

Keyword Propagation across a Complex Structure

❑ After keyword propagation

$$T' = T + P = T + TM = T(I + M) = TM_I$$

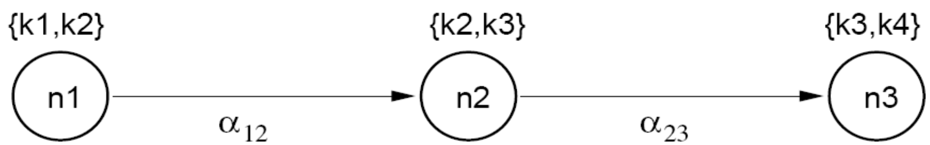
New *enriched* term-node matrix

Propagation Adjacency matrix

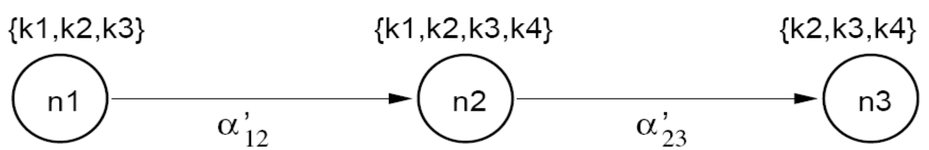
All diagonal values are 1 and all non-diagonal entries are same with M

Keyword Propagation across a Complex Structure

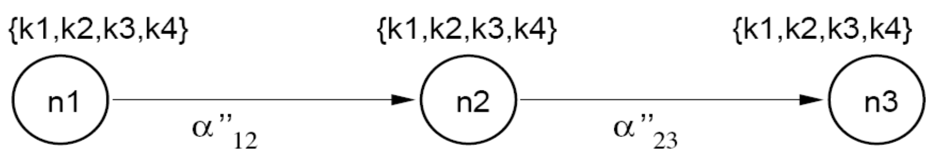
❑ Keyword Propagation Process



(a) Original content



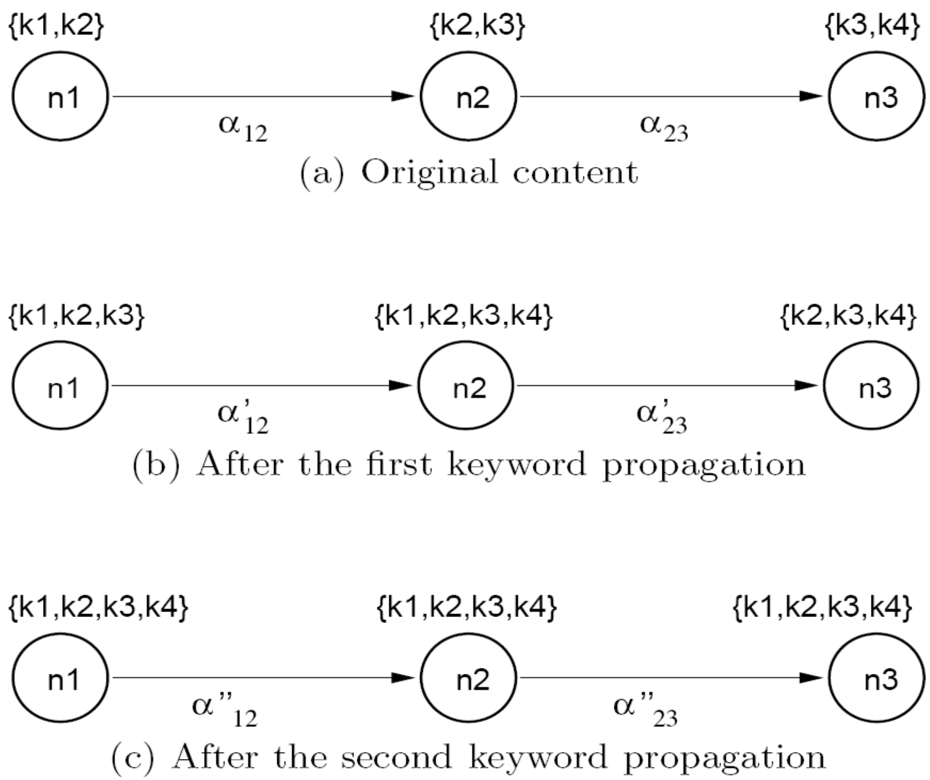
(b) After the first keyword propagation



(c) After the second keyword propagation

Keyword Propagation across a Complex Structure

❑ Keyword Propagation Process



(a) Original content

(b) After the first keyword propagation

(c) After the second keyword propagation

d = 2

$$T_{final} = T M_{l1} M_{l2} \ldots M_{ld}$$

Propagation adjacency matrix computed for the *d*th iteration

(d is the greatest number of edges between any nodes)

Table

- Motivation
- Approach
- Related Work
- Relative Content of Entries
- Keyword Propagation
    - Keyword Propagation between a Pair of Entries
    - Keyword Propagation across a Complex Structure
- Experiment
- Conclusion and Future Work

Experiment

❑ Experiment Setup

- ❑ Data
  - ☆ Yahoo Hierarchy
  - ☆ Computer Science, Mathematics, and Movie directory
- ❑ Ground truth and Query
  - ☆ 10 sample keyword queries
  - ☆ User study (8 users)

r	Relaxed	Differentiated	Strict
irrelevant	0	0	0
partially relevant	1	0.5	0
fully relevant	1	1	1

Experiment

❑ Experiment Setup

- ❑ Query processing
  - ☆ N (No Keyword Propagation)
  - ☆ KP (Keyword Propagation)
  - ☆ D_t and D_n
    - – No Keyword Propagation, but context extracted from the whole tree or neighbor
  - ☆ KP+ D_t and KP31+D_n
    - – keyword Propagation, and context extracted from the whole tree or neighbor

- ❑ Evaluation measure
  - ☆ P@10
  - ☆ MRR (Mean reciprocal rank of the first relevant document)
  - ☆ Paired t-Test

Keyword Propagation/ No Propagation

	N	KP	Improvement
Relaxed	0.670	0.753	12.27%
Differentiated	0.542	0.612	12.60%
Strict	0.415	0.469	13.10%

P@10

	N	KP	Improvement
Relaxed	0.869	0.930	6.97%
Differentiated	0.869	0.930	6.97%
Strict	0.644	0.730	13.20%

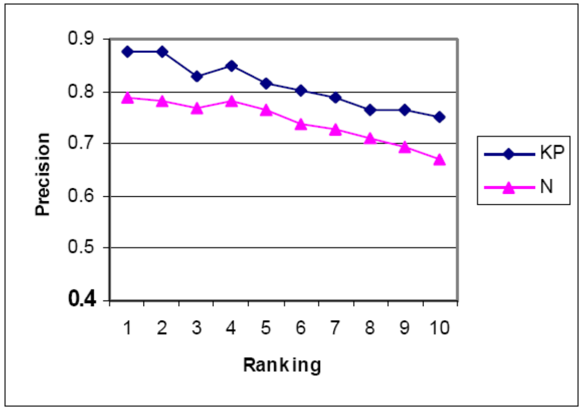
Average MRR

Keyword Propagation/ No Propagation

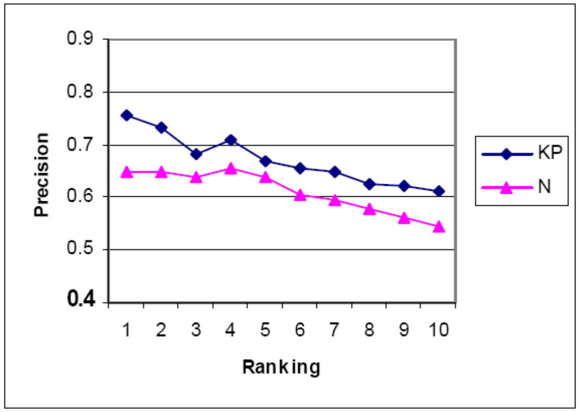
p-values for KP $vs.$ N	Relaxed	Differentiated	Strict
	0.029	0.031	0.047

P-values for the t-Test

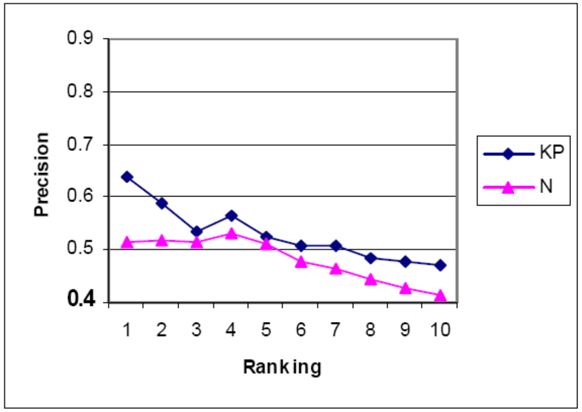
Keyword Propagation/ No Propagation



(a) Relaxed Precision vs. Ranking



(b) Differentiated Precision vs. Ranking



(c) Strict Precision vs. Ranking

Keyword Propagation/ Alternative Context Extraction

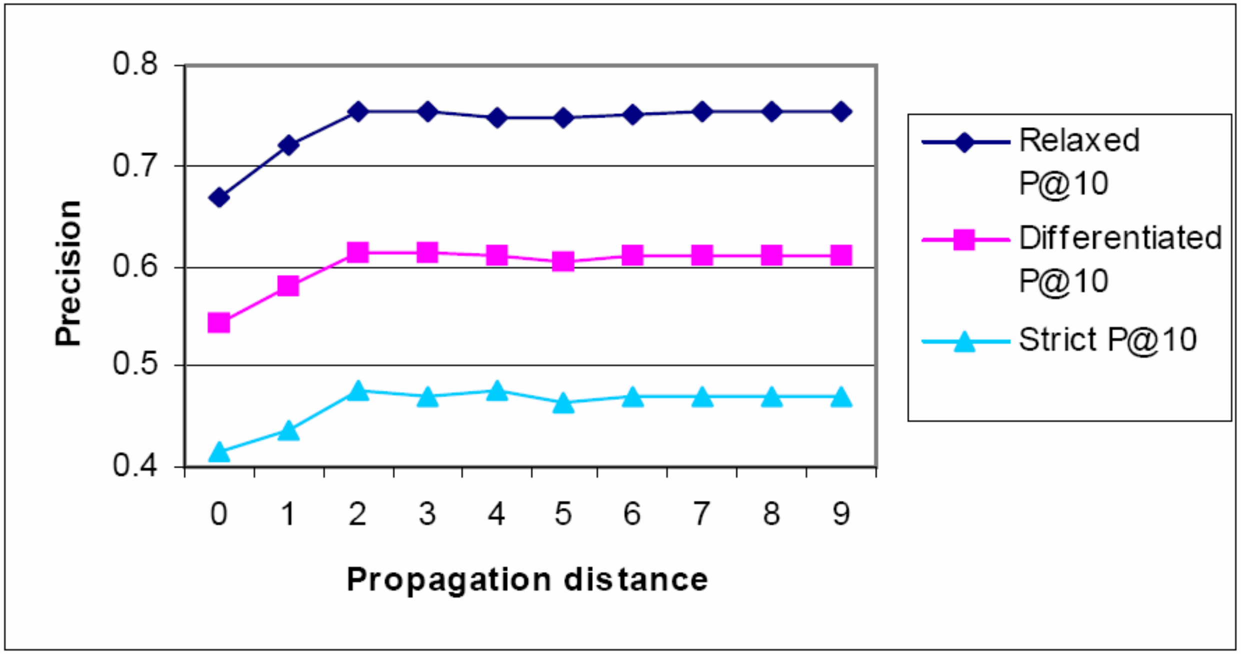
Differentiated; P@10						
β/γ	1/0	0.8/0.2	0.6/0.4	0.4/0.6	0.2/0.8	0/1
N	0.542	-	-	-	-	-
D_t	-	0.539	0.545	0.579	0.558	NA
D_n	-	0.532	0.542	0.547	0.564	0.572
KP	**0.612**	-	-	-	-	-
KP+D_t	-	0.606	0.607	0.607	0.597	NA
KP+D_n	-	0.611	0.612	0.596	0.584	0.572

Differentiated: P@10

Differentiated; t-Test						
β/γ	1/0	0.8/0.2	0.6/0.4	0.4/0.6	0.2/0.8	0/1
D_t vs. N	-	worse	55.1%	84.4%	63.5%	NA
D_n vs. N	-	worse	54.0%	65.2%	81.1%	90.5%
KP vs. N	**96.9%**	-	-	-	-	-
KP+$D_t vs$N	-	96.2%	95.7%	95.7%	90.0%	NA
KP+$D_n vs$N	-	96.7%	96.8%	91.8%	86.5%	90.5%

Differentiated: t-Test relative
No Keyword Propagation

Effect of the Structural Distance



Statistical Validation of the Ground Truth

- ANOVA test
  - A statistical test to observe the agreement between the assessors
  - We Identified two users whose judgments were significantly different from the other 6 users
  - When excluding these two users, the user judgments were in agreement

Statistical Validation of the Ground Truth

Differentiated; P@10						
β/γ	1/0	0.8/0.2	0.6/0.4	0.4/0.6	0.2/0.8	0/1
N	0.538	-	-	-	-	-
D_t	-	0.547	0.556	0.594	0.571	NA
D_n	-	0.525	0.537	0.544	0.565	0.573
KP	**0.628**	-	-	-	-	-
KP+D_t	-	0.625	0.624	0.624	0.608	NA
KP+D_n	-	0.625	0.625	0.614	0.601	0.573

Differentiated: P@10

Differentiated; t-Test						
β/γ	1/0	0.8/0.2	0.6/0.4	0.4/0.6	0.2/0.8	0/1
D_t vs. N	-	62.0%	71.3%	80.4%	72.1%	NA
D_n vs. N	-	worse	worse	69.9%	74.9%	91.8%
KP vs. N	**97.3%**	-	-	-	-	-
KP+$D_t vs$N	-	96.4%	96.6%	96.6%	90.5%	NA
KP+$D_n vs$N	-	96.4%	96.4%	93.8%	90.5%	91.8%

Differentiated: t-Test relative
No Keyword Propagation

Table

- ❑ Motivation
- ❑ Approach
- ❑ Related Work
- ❑ Relative Content of Entries
- ❑ Keyword Propagation
  - ❑ Keyword Propagation between a Pair of Entries
  - ❑ Keyword Propagation across a Complex Structure
- ❑ Experiment
- ➢ Conclusion and Future Work

Conclusion and Future Work

- Conclusion
  - Present a technique to identify a semantic relationship
  - Introduce a relative content preserving keyword propagation technique
- Future Work
  - Incorporate of other types of semantic cues
    - ☆ Structured-based method
    - ☆ Information-based method

Question